ICIC Express Letters
Part B: Applications
Volume **7**, Number **6**, June **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **1179–1184**

REFINEMENT OF FACE ALIGNMENT BY CASCADED SHAPE REGRESSION

Jun Kong[1,2], Tao Yang[1], Min Jiang[1], Xiaofeng Gu[1] and Shengwei Tian[2]

[1]Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education)
Jiangnan University
No. 1800, Lihu Avenue, Wuxi 214122, P. R. China
kongjun@jiangnan.edu.cn

[2]College of Electrical Engineering
Xinjiang University
No. 1230, Yan'an Road, Urumqi 830047, P. R. China
xuanyuyt@163.com

Received November 2015; accepted February 2016

Abstract. *This paper presents a method of an improvement in cascaded shape regression for face alignment. Firstly, our method begins with detecting several key points so as to coarsely locate the initial shape. Secondly, a local region with radius refines the large feature pool for feature selection. Lastly, we eliminate regressor which cannot minimize the alignment errors during the training. The main contributions of our work include: i) preventing final iteration from being trapped in local optima due to the poor initial shape, ii) efficiently avoiding over-fitting caused by extremely large feature pool that brings about noise features in feature selecting, and iii) reducing the number of weak regressors to enhance regression efficiency. Extensive experiments on COFW and HELEN face datasets demonstrate that our method outperforms the traditional cascaded regression methods.*
Keywords: Cascaded, Face alignment, Feature pool, Refining, Initial shape

1. **Introduction.** Given a face image, face alignment is achieved by estimating a shape S that consists of M facial landmarks, making S as close as possible to the target shape $\hat{S}$, i.e., minimizing the alignment error

$$S = \arg\min_{S} \left\| \hat{S} - S \right\|_2 \tag{1}$$

Among many different methods for face alignment, cascaded shape regression [1-4] has emerged as the leading and state-of-the-art method. Explicit Shape Regression (ESR) [1] and Robust Cascaded Pose Regression (RCPR) [2] use boosted regression and random $fern$ [3] to regress selected discriminative pixel-difference features. Supervised Descent Method (SDM) [4] is proposed for minimizing a Non-linear Least Squares (NLS) function. Local Binary Features (LBF) [5] uses random forest to encode the local binary features for each landmark independently.

These algorithms typically start from an initial shape S^0, e.g., mean shape [2, 5] of training samples, and progressively refine the shape estimations to output final shape estimation S. In practice, the initial shape may be far from the target shape. So the discrepancy between them is unlikely rectified by estimating a shape increment ΔS stage-by-stage. As a consequence, the alignment may be trapped in local optima. Both [1, 2] use a *fern* as a weak regressor. Randomly sampling P pixels, in total P^2 pixel-difference features are generated. Out of P^2 features, they chose F pixel-difference features. This converts into unaffordable task if they want to learn the most efficient feature combination. In particular, there are many noise features in this large feature pool, which can easily cause over-fitting and hurt performance in testing.

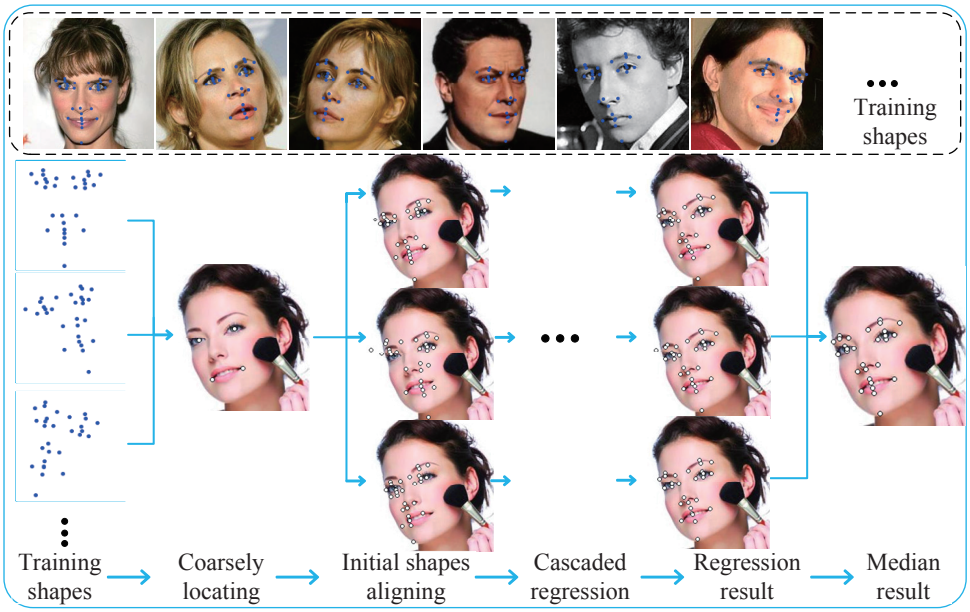

FIGURE 1. Coarse-to-fine cascaded shape regression

Aiming at the shortcomings of the traditional cascaded shape regression, we propose three improvements on it. Firstly, inspired by the fact that only a few salient landmarks that can be reliably characterized in image, we coarsely locate eye centers and mouth corners to help initial shape approach target shape, as shown in Figure 1. Secondly, we refined feature pool to ensure to eliminate noise features and simultaneously preserve the most discriminative texture information. Finally, since each weak regressor needs to slightly minimize the sum of alignment errors, we eliminate regressor which cannot reduce errors during the training.

The paper is organized as follows. Section 2 presents our face alignment framework. Section 3 conducts experiments to demonstrate the performance of our method. Section 4 places the conclusion.

2. Revised Shape Regression.

In this section, we detail our method on initial shape coarse alignment, feature pool refining and regressors screening.

2.1. Key points detection and initial shape coarse alignment.

The cascaded regression approach begins with an initial shape S^0 that comes from training samples, and refines the shape through sequentially trained regressors. While the initial shape is far from the target shape, the alignment process may be trapped in local optima.

In order to obtain the superior initial shape, we firstly use the Structured-Output Regression Forests (SO-RF) [6] algorithm to coarsely detect the facial key points of eye centers and mouth corners. The key to SO-RF lies on incorporating the structure information of facial features, within the regression forests framework.

Then, a similarity transformation can be got by the initial shape and facial key points. By the similarity transformation, we can obtain the center, rotation and scale of the initial shape to target shape. We use multiple initial shapes by augmenting training data to estimate for one face image. At last, we take the median of all estimation as final output. This data augmentation method also has been adopted in [2, 5].

2.2. Cascaded regression.

Cascaded regression is formed by a cascaded of T weak regressors $(R^1, R^2, \ldots, R^T)$. At each stage, regressors R^t produce an update ΔS^t, which is then combined with previous iterations estimate S^{t-1} to update shape estimation S^t.

$$S^t = S^{t-1} + R^t\left(I, \Delta\hat{S}^t\right), \quad t = 1, \ldots, T \tag{2}$$

where I is originally image and $\Delta\hat{S}^t = \hat{S} - S^{t-1}$ is regression target of each stage.

To measure alignment errors between two shapes, we require a function:

$$err\left(S^a, S^b\right) = \left\|S^a - S^b\right\| \tag{3}$$

Using N training samples $\left\{\left(I_i, \hat{S}_i, S^0\right)\right\}_{i=1}^{N}$, each regressor R^t is trained to minimize the sum of alignment errors:

$$R^t = \arg\min_R \sum\nolimits_{i=1}^{N} err\left(\hat{S}_i, S_i^t\right) \tag{4}$$

We compare alignment errors before and after iteration to decide whether regression is convergent or not:

$$\epsilon_t = \sum\nolimits_{i=1}^{N} err\left(\hat{S}_i, S_i^t\right) \Big/ \sum\nolimits_{i=1}^{N} err\left(\hat{S}_i, S_i^{t-1}\right) \tag{5}$$

If $\epsilon_t \geq \theta$ ($\theta \leq 1$), training stops; otherwise, we continue training. After training R^t, we apply Equation (2) to update S^t for the next phase of training.

Each regressor R^t is a boosted regressor, e.g., $R^t = \left(r^{t,1}, r^{t,2}, \ldots, r^{t,K}\right)$. Moreover, since each weak regressor $r^{t,k}$ needs to slightly minimize the sum of alignment errors, we define the relative errors of each weak regressor $r^{t,k}$ as:

$$\epsilon_{t,k} = \sum\nolimits_{i=1}^{N} err\left(\hat{S}_i, S_i^{t-1} + r^{t,k}\left(I_i, \Delta\hat{S}_i^t\right)\right) \Big/ \sum\nolimits_{i=1}^{N} err\left(\hat{S}_i, S_i^{t-1}\right) \tag{6}$$

When $\epsilon_{t,k} \leq 1$, regressor $r^{t,k}$ is reserved; otherwise, we abandon it. Algorithm 1 shows the main steps of the cascaded regression procedure. Finally, we will get K' ($K' \leq K$) weak regressors r. These regressors selecting mechanism, make each weak regressor slightly minimize the alignment errors, and efficiently improve the efficiency of regression.

Algorithm 1 Training for cascaded shape regression

Input: training data $\left\{\left(I_i, \hat{S}_i, S^0\right)\right\}_{i=1}^{N}$

for $t = 1 \rightarrow T$ **do**
  $\Delta S^t \leftarrow 0$, $\Delta\hat{S}^t \leftarrow \hat{S} - S^{t-1}$
  for $k = 1 \rightarrow K$ **do**
    Compute a fern output $r^{t,k}\left(I, \Delta\hat{S}^t\right)$
    if $\epsilon_{t,k} \leq 1$ **then**
      Update $\Delta S^t \leftarrow \Delta S^t + r^{t,k}\left(I, \Delta\hat{S}^t\right)$, $\Delta\hat{S}^t \leftarrow \Delta\hat{S}^t - r^{t,k}\left(I, \Delta\hat{S}^t\right)$
    end if
  end for
  Update $S^t \leftarrow S^{t-1} + \Delta S^t$
  if $\epsilon_t \geq \theta$ **then**
    stop training
  end if
end for

Output: $R = \left\{r^{1,1}, r^{1,2}, \ldots, r^{t,k}, \ldots\right\}$

2.3. **Feature pool refining.** Encouraged by the success of random $fern$ for classification [3], we use a standard $fern$ as each regressor r. Our features use simple pixel-difference features [5, 7]. We randomly sample P pixels and index a pixel p by its local coordinates $(\delta x, \delta y)$ with respect to nearest landmark m of S^{t-1}.

Using the entire face region as the sample region will result in many noise features or poor discriminative features in the large feature pool. In our work, we refine this large feature pool to select the most discriminative feature combination.

Considering the most discriminative texture feature lies in a local region around the estimated landmark S^{t-1}, we eliminate pixels that are far from previous estimated landmark from a local region with radius ω. As an overview of the whole approach, we list the major refining mechanism in Algorithm 2. The process will finally generate L^2 index features. Now, the new challenge is how to get the optimal radius ω.

Algorithm 2 Obtaining and refining the feature pool

Input: radius ω and normalized mean shape $\bar{S} = \{\bar{x}_1, \bar{y}_1, \ldots, \bar{x}_M, \bar{y}_M\}^T$, $\bar{x}, \bar{y} \in [-1\ 1]$

for $i = 0 \rightarrow P$ **do**
  Randomly sample a point $p_i(x_i, y_i)$ from the $[-1\ 1]$ region
  Set $d_0 \leftarrow \infty$
  for $j = 1 \rightarrow M$ **do**
    $d_j \leftarrow \sqrt{(x_i - \bar{x}_j)^2 + (y_i - \bar{y}_j)^2}$
    if $d_j < d_{j-1}$ **then**
      $d_{\min} \leftarrow d_j$, $m \leftarrow j$, $\delta x \leftarrow x_i - \bar{x}_j$, $\delta y \leftarrow y_i - \bar{y}_j$
    end if
  end for
  if $d_{\min} < \omega$ **then**
    Collect index features f with index m and local coordinate $(\delta x, \delta y)$
  end if
end for

Output: a vector index features $\{f_1, f_2, \ldots, f_L\}$ $\quad (L \leq P)$

We believe that this radius ω is related to each stage distribution of $\Delta\hat{S}^t$. For efficiently computation, we use simple the average $\Delta\hat{S}$ of all landmarks. To get the optimal radius ω, we train different models with various radii, and testing samples with test errors. We repeat the experiment and take the radius with minimum test error as the best radius ω. Our experiment demonstrates that the optimal radius will diminish stage-by-stage in iteration. If $\Delta\hat{S}$ of all training images distribute widely, the best radius ω is a big one; otherwise, it is a small one.

Finally, we select F pixel-difference features to construct a $fern$ as [1, 2] proposed.

3. **Experiments.** Our implementation of the proposed method is based on the Explicit Shape Regression code provided by [1]. We empirically set $T = 7$, $K = 500$, $P = 400$, $F = 5$, and $\theta = 0.91$, and the parameters of ω are shown in Table 1. We provide comparable results with other methods on the *COFW* [2] and *HELEN* [8].

TABLE 1. The number of weak regressors K', the optimal local region radius ω and corresponding L at each stage for *COFW* dataset

stage	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6	Stage 7
K'	487	417	376	301	277	212	136
ω	0.34	0.21	0.15	0.11	0.09	0.078	0.067
L	396	370	340	323	289	286	280

3.1. **Performance evaluation.** Our competitors are the shape regression based methods, including ESR and LBF. The performance is measured on Intel Core i5 3.20GHz CPU with Ubuntu C++ implementation. The overall accuracy is reported based on the Averaged Error (AE) and Cumulative Error Distribution (CED) curve to cater for different evaluation schemes in the literature. We consider any errors above 10% to be a failure, as suggested in [9]. Overall, Figure 2 presents some examples and comparable results.

FIGURE 2. Example images from the *COFW* dataset where our method outperforms ESR and LBF

TABLE 2. Averaged error (AE) and failure rates

COFW (29 landmarks)			***HELEN*** (68 landmarks)		
Method	AE	failures	Method	AE	failures
ESR	9.63	33%	ESR	5.51	3.6%
LBF	8.34	20%	LBF	5.37	6.2%
Our Method	**6.37**	**11%**	Our Method	**5.33**	**3.6%**

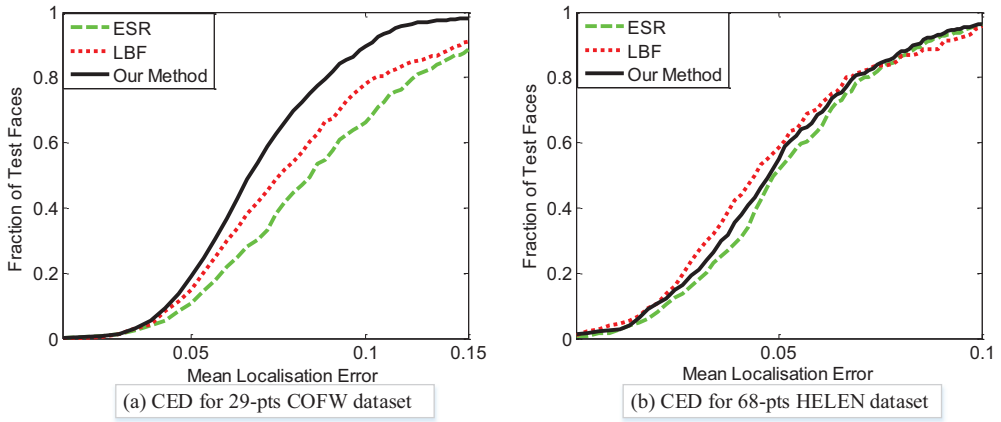

FIGURE 3. Comparisons of cumulative errors distribution (CED) curves

Averaged error comparison: Table 2 is the comparison of our algorithm with other two algorithms about AE and failure rates in various datasets. The errors are normalized by dividing with the interocular distance. It can be observed if large variations of head pose and occlusion occur in the datasets, our method outperforms other two methods; otherwise it is on par with the LBF. The results demonstrate that the robustness of the feature pool refining and the initial shape coarse alignment framework outperform the conventional cascaded regression methods.

Cumulative error distribution comparison Figure 3 is the comparison of our method with other two algorithms about CED curves. Again, CED mainly depends on large head pose and partial occlusion. In addition, our method is more efficient to overcome large head pose and partial occlusion, while LBF is more accurate in a simple environment with no large variations of head pose or occlusion.

3.2. **Further analyses.** In principle, the final selected F features cannot provide sufficient information in the entire face region, which leads to reducing accuracy. However, the

feature pool refining and the initial shape coarse alignment mechanism make our method efficient in eliminating many noise features and overcoming large head pose. The feature pool refining provides a feasible way to select the most discrimination feature combination quickly and accurately.

The feature pool refining plays an important role in our method in offering a better feature pool for selecting the most discriminative feature combination. For *COFW* dataset, Table 1 shows the number of weak regressors K', the optimal local region radius ω and correspondingly L in every stage. After feature pool refining, the size of feature pool is decreased from P to L. The refining mechanism eliminates much noise and reduces training costs especially in the later stages. Without this process, the averaged error increases to 7.5. Furthermore, the initial shape coarse alignment is introduced into our method to overcome large head pose, especially in the early stages. If we simply put initial shapes from a fixed range of neighbourhood of mean shape, the averaged errors would increase from 6.37 to 8.5. Errors are mainly observed in cases of large head pose. Finally, the regressors screening mechanism, also makes the number of regressors decrease from K to K', and enhances regression efficiency.

4. **Conclusion.** Occlusions and large head pose are two main challenges for current face alignment methods. To handle the challenges mentioned above, in this paper, a novel method based on initial shape coarse alignment, feature pool refining and regressors screening is proposed. Our method could prominently improve the robustness of traditional cascaded regression methods. The initial shape coarse alignment and feature pool refining can efficiently locate the landmarks with large head pose and eliminate noise features before selecting the most discriminative feature combination respectively. Last, our regressors screening mechanism makes regression more efficient. In the future, we plan to encode the local texture for landmarks to learn intrinsic features for more effective regression.

Acknowledgment. This work is partially supported by Xinjiang Uygur Autonomous Regions University Science and Research Key Project (XJEDU2012I08), National Natural Science Foundation of China (61362030, 61201429), the Project Funded by China Postdoctoral Science Foundation (2015M581720), Technology Research Project of The Ministry of Public Security of China (2014JSYJB007).

REFERENCES

[1] X. Cao, Y. Wei, F. Wen and J. Sun, Face alignment by explicit shape regression, *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp.2887-2894, 2012.

[2] X. P. Burgos-Artizzu, P. Perona and P. Dollar, Robust face landmark estimation under occlusion, *2013 IEEE International Conference on Computer Vision*, pp.1513-1520, 2013.

[3] M. Ozuysal, M. Calonder, V. Lepetit and P. Fua, Fast keypoint recognition using random ferns, *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp.448-461, 2010.

[4] X. Xiong and F. de la Torre, Supervised descent method and its applications to face alignment, *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp.532-539, 2013.

[5] S. Ren, X. Cao, Y. Wei and J. Sun, Face alignment at 3000 FPS via regressing local binary features, *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp.1685-1692, 2014.

[6] H. Yang and I. Patras, Face parts localization using structured-output regression forests, *Computer Vision*, Springer, pp.667-679, 2013.

[7] Z. Li, X. Liu, X. Duan and C. Wang, An efficient detection approach for facial geometric features, *ICIC Express Letters*, vol.7, no.10, pp.2837-2842, 2013.

[8] V. Le, J. Brandt, Z. Lin, L. Bourdev and T. S. Huang, Interactive facial feature localization, *Computer Vision*, Springer, pp.679-692, 2012.

[9] M. Dantone, J. Gall, G. Fanelli and L. Van Gool, Real-time facial feature detection using conditional regression forests, *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp.2578-2585, 2012.